Uber

An Experiment Proposal On The Effect Of Uber's Treating Its Drivers As Employees Rather Than Independent Contractors

By Shivkumar Umadi, Hing Kit Leung, Jeeyoung Lim, ShengYa Mei

Overview

Currently, Uber's drivers are working as independent contractors. Uber claims that this employment status brings flexibility and independence to drivers and provides valuable experience for customers.

However, these days, Uber faces challenges regarding drivers' employment. In multiple states, there have been several lawsuit cases where Uber drivers argued that Uber has not truly given them independence and tried to avoid giving drivers the benefits and protection of employment. Furthermore, some drivers are trying to legalize changing their employment status to regular employees and the UK supreme court has ruled that Uber should hire drivers as regular workers. This might have an enormous impact on Uber's business. It could increase costs but also bring positive results as well. So, we would like to measure the risk and benefits of the policy change through this experiment for important decision-making. In this experiment, we are going to look into the front-end part of the business, such as the drivers and customers.

Setting The Stage

Objective Function and Strategic priorities:

We want to know whether the pending regulation would have an adverse impact on the services we provide to the drivers and riders, and if so, what is the magnitude of the impact. The objective of the experiment is to check the effect of the implementation of the law that the government has stated, that employees working for an organization have to be full-time employees with specific rules in place. Additionally, we need to also understand the effect of the law on Uber drivers becoming full-time employees.

The strategic priority of this experiment on Uber drivers is not to implement the entire law in one go. We need to check the effect of the change stepwise. Hence, we consider 3 treatment groups based on the level of change implemented and understand its effect on Uber drivers and then scale it to a higher level of treatment. With this, we do not perform the change all at once, rather we do it gradually to better understand the effect of this new implementation.

Treatments:

To reduce complexity, we assume that the law, if effective, only requires that Uber provide minimum wage, insurance, and pension. As a result, we want to implement policies that can counteract the excessive costs, so the treatments, we think, are the ones that may result in certain benefits to the company at the expense of drivers' flexibility. The set of treatment variations, ordered in decreasing flexibility, for the drivers are 1) Wearing a uniform when on duty, 2) Required to receive employees' training sessions and adhere to certain customer service protocols, and 3) Fixed working hours.

This experiment takes the form of a mechanism test because we are introducing a new policy to Uber's ride-sharing services and underlying policy change has an overarching effect on the

way Uber runs the business. Through this experiment, we seek to understand the direction of the new policy effects which can go both ways.

Design Of Outcome Metrics:

Rather than one, we came up with 6 outcome metrics to measure Drivers' and Riders' perceptions of the change in the policy. In terms of drivers, we have to know the number of drivers who quit, their average job satisfaction, and the number of new driver applications, while for riders, metrics, such as their riding satisfaction, the tips they give per ride, and the number of new Uber app users.

Besides the riding/job satisfaction, the rest of the metrics are the proxies for the measurement of our objectives. Average Tips/bill, for example, is somewhat reflective of the riding experience and the service we provide. The higher it is, we expect the more positive the treatment effect is.

To be honest, deciding on time is a tricky business. We want to keep it as short as possible to lower costs while making it as long as necessary to sufficiently capture the treatment effects. After a thorough discussion with teammates, we concluded that the experiment is conducted over 6 months. The reasons behind the decision are that we want to give sufficient time for the news to sink in among drivers and riders and that we hope to leave ample time to devise and execute strategies based on the data we retrieve.

The fact that the metrics are measured in different scales requires us to normalize the data before assembling them into an Overall Evaluation Criteria (OEC) metric. The formula we adopt to perform normalization is: $$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$,

in which Xnorm is the normalized score, X is a score of a metric, and min(x) and max(x) are the minimum and maximum scores respectively, resulting in all scores ranging from 0 to 1 across metrics. By default, the weightings among the metrics are equally divided with the number of drivers quit being negative. But we do expect a discussion in which field experts provide their insights on how to better weight the metrics.

Considering the costs of implementing the policy if the bill was passed, we expect a large effect size even in the drivers' and riders' aspects alone. Treatment would be taken up if the OEC metric is 20% greater than the control group, and conversely, we anticipate an increase in the spending to lobby against the law in the U.S. congress if it is 20% less. If neither, we may think of new treatments to counter or mitigate the costs incurred when the law is imposed in the next round of the experiment.

Scaling up Effects:

As we see, Uber operates in 267 cities in the US. In our experiment, we consider 50 cities for control and 150 cities for the 3 treatment groups. We choose to consider 200 cities for the experiment. Hence, considering the entire population of 267 cities should not cause a problem.

Secondly, we choose the population based on an individual's income level to segregate the sample size for each of the treatment and control groups. Accordingly, the population left out (in this case: 67 cities) would fall in one of the stratification cases. Hence, the direction of our experiment seems appropriate for scaling it to a bigger population.

Uber

Implementation Planning

Experiment Design

The target population of our experiment is Uber drivers and **Uber ride-sharing customers**. These populations are the right target selection because the experiment aims to test the effect of a policy change on the front-end part of the business. In designing the experiment, the Uber driver population will be divided into 1 control group and 3 treatment groups. Drivers in the control group will receive the current Uber policy. They will remain unchanged in their function as freelancing employees/ independent contractors who have the option of working in other ride-sharing agencies.

Control - Drivers abide by the current Uber policy and function as independent contractors with no restrictions on external ride-sharing companies' orders

Drivers in each of the 3 treatment groups will function as full-time employees of Uber. All Uber drivers in treatment groups will receive minimum wage and are prohibited from receiving orders from external ride-sharing companies. In addition, each treatment group will abide by a new uber policy with the following distinctions:

Treatment 1 - Drivers must be in uniform when on duty
Treatment 2 - Drivers must receive employees training and adheres to certain customer service protocols
Treatment 3 - Drivers have fixed working hours. Drivers can specify their desired hours on a given day but have to meet a minimum of 6 hours and a maximum of 8 hours.

Currently, Uber operates across 267 cities in the United States (*uber.com*). Considering the majority of cities, 200 US cities will be randomly selected to participate in the experiment with 50 cities in each of the control and 3 treatment groups. Drivers will be randomly clustered on the city level based on the group they are in. Half of the Uber driver population in each city will be selected to participate in the experiment (*The choice of driver population by city explained later*). Population size is detailed as follows:

Total number of US cities - **200**
Cities in Control - **50**
Cities in Treatment 1 - **50**
Cities in Treatment 2 - **50**
Cities in Treatment 3 - **50**
Driver count in each city - **50% of the uber driver population in that city**

To account for major imbalances in our outcome metrics and treatment effects (e.g. tips), cities will be stratified into 3 income levels detailed as follows:

1. High-tier if average income >= 100K
2. Mid-tier if average income >= 50K and <100K
3. Low-tier if average income < 50K

*We assume high-average-income cities will give higher tips.

Uber drivers are allowed to operate across cities. To avoid interference among treatment groups and to capture income stratification among cities, we will update the Uber App such

that customers requesting a within-city ride, will be paired with a driver participating in the experiment. This applies to all treatment groups including control. Customers requesting a cross-city ride will be paired with drivers not participating in the experiment. We will select **50%** of the driver population in each city to participate in the experiment such that Uber can still operate their ride-sharing service to standard when customers request cross-city rides.

There still exists the possibility that drivers in treatment groups will interfere with drivers not participating in the experiment. Considering that this exposure is equally distributed among treatment groups and that we are not collecting data from non-participants, we suspect the effect to be minimal when assessing the integrity of our outcome metrics.

In initiating the experiment, announcements regarding policy changes will be made to each driver based on the treatment group they are in. Drivers will receive the announcement 2 weeks in advance. We think 2 weeks' notice During this phase, there is a likelihood of fluctuation in sample size. Potential loss or gain of Uber drivers should be made aware of before the experiment goes into effect. The actual experiment will go on for 6 months to allow for a comprehensive evaluation.

Collection of Outcome Measures:

In total, we have 6 outcome metrics, 3 for customers and 3 for drivers'. In line with the way, we randomize our sample, the unit of analysis will be conducted in clusters, namely cities in this case. However, the method by which we collect and transform the data differs among the metrics. On one hand, riders' satisfaction, tips per ride, and drivers' satisfaction will be collected first at the individual level and then calculated on average by each city. On the other hand, metrics, such as the number of new Uber users, of drivers who quit, and driver applications, are recorded over the 6-month experimentation period for each city.

The data of most of the metrics, according to our knowledge, can be readily retrieved through Uber's data warehouse. Riding satisfaction, for example, is the equivalent of the survey that popped up in the app and asked riders to rate the experience on a scale of 1 to 5 after taking the Uber. Conversely, we can reasonably assume that Uber has never issued surveys to drivers asking about job satisfaction because they are independent contractors who leverage Uber's tech to make earnings. A different survey asking "On a scale of 1 to 5, how satisfied are you working with Uber", will be sent out to drivers in the control and treatment group twice a month. This rate of distribution is deliberately decided so that we cause the least amount of disturbance to the drivers while factoring in their emotional fluctuation from time to time.

Unlike natural experiments, such as event studies and differences in differences, randomized experiments, like ours, do not emphasize the time variable. The outcome metrics in the control are served as the baseline against which those in the treatment compare. As for the risk of attrition, it is not relevant in our case. Individuals' dropping out of our study is expected and measured, and hence it won't affect our ability to interpret and make use of the results.

Sample Size

As we try to measure the drivers' and riders' perceptions of the change in our policy, which is very subjective, this underlying variability in the outcome measures is expected to be great. Lack of data, we can only roughly estimate the variability, namely the sigma, to be 0.2. As stated in the previous part, the minimum detectable effect size should be 20% less or greater than the control mean so that the treatment is considered different from the control.

In terms of determining the “Significance Threshold”, we understand that it, in a statistical sense, means the probability of falsely rejecting the null hypothesis, which in business, results in us taking futile actions. Given the costs of us mistakenly thinking the treatment, if the law is effective, can counter the costs, we need to lower the “Significance Threshold” to 0.001.

Another parameter that decides our sample size is “Power”, which can be interpreted, in a statistical sense, as the probability of correctly rejecting the null hypothesis. In business, we consider it as the level at which we are confident that we take up the treatment that is, in fact, effective to achieve our objective function. Since the consequence of failing to implement the necessary actions, increase spending in lobbying the U.S. Congress, is grave, we set the “Power” as 0.96

To put the above data together, the minimal sample size is 200 in this 2-sided test. As of today, Uber operates its share-drive business in nearly 267 U.S. cities. Given this understanding, time and budget would not be a major constraint in collecting a sufficient sample size.

Test Selection:

OEC from each treatment group will be used to compute the average treatment effect (ATE) between each of the treatment groups and the control group. We will then use the ANOVA test to verify the significance of means, in our case OECs, among treatment and control groups. If the ANOVA test on means came out significant between treatment 1 and control, for example, we would conclude that their ATE is statistically significant.